DOCUMENT RESUME

ED 400 282 TM 025 550

AUTHOR Gershon, Richard C.; And Others
TITLE Analyzing Multiple Choice Tests with the Rasch Model: Improving Item Calibrations by Deleting Person-Item Mismatches.
PUB DATE Apr 94
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adaptive Testing; Computer Assisted Testing; Difficulty Level; Elementary Education; *Estimation (Mathematics); Goodness of Fit; Item Banks; *Item Response Theory; *Multiple Choice Tests
IDENTIFIERS Calibration; *Rasch Model

ABSTRACT

A 1992 study by R. Gershon found discrepancies when comparing the theoretical Rasch item characteristic curve with the average empirical curve for 1,304 vocabulary items administered to 7,711 students. When person-item mismatches were deleted (for any person-item interaction where the ability of the person was much higher or much lower than the difficulty of the item), the difference between the theoretical and empirically produced curves was decreased. This paper describes a replication of that original study, using data obtained from an administration of the California Achievement Test to students in the Minneapolis (Minnesota) public schools. When person-item mismatches were deleted, item calibrations improved regardless of the grade level. The results are discussed with an emphasis on the importance of selectively deleting data when the primary goal of the analysis is to obtain the most accurate item difficulty estimates possible. This research is of particular importance for testing organizations that use item banks or computerized adaptive testing. (Contains one table, six figures, and eight references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
BETTY BERGSTROM

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 400 282

Analyzing Multiple Choice Tests with the Rasch Model: Improving Item Calibrations by Deleting Person-Item Mismatches

Richard C. Gershon
Betty A. Bergstrom

Computer Adaptive Technologies, Inc.

William L. Brown
Minneapolis Public Schools

Paper presented at the annual meeting
American Educational Research Association

New Orleans, Louisiana

April, 1994

TM 025550

BEST COPY AVAILABLE

Abstract

A 1992 study by Gershon found discrepancies when comparing the theoretical Rasch Item Characteristic Curve with the average empirical curve for 1,304 vocabulary items administered to 7,711 students. When person-item mismatches were deleted (for any person item interaction where the ability of the person was much higher or much lower than the difficulty of the item), the difference between the theoretical and empirically produced curves was decreased. This paper describes a replication of that original study, using data obtained from an administration of the California Achievement Test to students in the Minneapolis Public Schools. When person-item mismatches were deleted, item calibrations improved regardless of the grade level. The results are discussed with an emphasis on the importance of selectively deleting data when the primary goal of the analysis is to obtain the most accurate item difficulty estimates possible. This research is of particular importance for testing organizations who use item banks and/or computerized adaptive testing.

Analyzing Multiple Choice Tests with the Rasch Model: Improving Item Calibrations by Deleting Person-Item Mismatches

Ideally, tests should be constructed to test examinees at their ability level. Yet, in many testing situations, some of the items on a test are inappropriately easy or difficult for some examinees. This situation may arise when both the test items and the test population are very heterogeneous, resulting in very able examinees responding to very easy items and vice versa. It may also occur when the purpose of the test is to show mastery. In this case many examinees may be more able than items are difficult.

A primary assumption of any assessment model is that only relevant variables are being measured. David Andrich (1989) pointed out that person ability can theoretically span an infinite range but that a given item should be expected to "operate consistently only in a specific range of the variable". Andrich explains, "The range within which a statement (item) may be expected to operate consistently can be constrained by considering the probability of a positive or negative response: if this is very high (say greater than 80%. . .) then the responses should be eliminated....The responses at the extremes simply should not be expected to work accurately." In practice, we often remove misfitting items before final test analysis. For this paper, our suggestion is not that entire items or persons be removed from the analysis, but rather that the data be edited to eliminate extreme item-person mismatches.

Previous research (Gershon, 1992) has shown that for vocabulary test items, when persons with ability far below that of a given item difficulty were included in the sample, guessing played a role in estimating the difficulty of the item. When the observed performance of each person-item interaction was compared to predicted performance for this sample,

discrepancies were noted in the observed item characteristic curve, particularly at the lower end of the scale. When the data set was edited, (person-item interactions were marked as missing when the ability of the examinee was more than 2 logits higher or 1 logit lower than the difficulty of the item) the discrepancy between the theoretical curve and the empirically unedited curve was minimized.

Other research has shown that when inappropriate samples of examinees are used for item calibration, the standard deviation of the item calibrations is constrained (Bergstrom, Gershon and Brown, 1993). This constriction results in a decrease in the range of item difficulty calibrations obtained, and ultimately in poor targeting of the items which were found to be particularly "easy" or "difficult" for the specific sample.

This paper explores the effect of poorly targeted tests on item parameter estimation. We examine the results of marking as missing, item-person interactions where the difficulty of the item varies considerably from the ability of the person. We replicate the original Gershon (1992) research, which included only vocabulary items, with a more general data set. To our knowledge, no additional research has been published or presented on systematically eliminating person-item mismatches from multiple choice data.

Method

Subjects

Data are from the Minneapolis Public Schools 1993 administration of the California Achievement Test. The California Achievement Test was selected by the Minneapolis Public Schools over other nationally normed tests because it most closely matched the district

curriculum (Brown, O'Gorman, Rogers and Alm, 1993). Four subtests of the Form E 1985 standard edition of the California Achievement Test were administered district wide in the spring of 1993 to grades 1,2,3,4,6,8 and 10 in vocabulary, reading comprehension, math computation, and math concepts and applications.

We independently analyzed each grade. The results of the analyses were very similar, and thus we report only the grade 1 results in this paper.

Method

The data were analyzed using BIGSTEPS (Wright and Linacre, 1992) under the following three conditions:

In Condition 1 there was no restriction of data (persons or items).

In Condition 2, we employed a recently added feature in BIGSTEPS, "CUTHI/LO" that allows the data set to be edited according to the specifications of the user.[1] "CUTHI" was set to 2 and "CUTLO" was set to 1. These parameters instructed the program to (a) estimate all person ability and item difficulty parameters using PROX; (b) examine each person-item interaction and mark as missing all person item interactions where the person ability estimate is 2 logits greater than the item difficulty estimate and all person item interactions where the person ability estimate is 1 logit lower than the item difficulty estimate; and (c) re-estimate item difficulties and person abilities using PROX and UCON iterations on the restricted data set.

Since most users would not accept person ability measures based on a restricted data set, we did an additional BIGSTEPS run in Condition 2 in which we anchored all item calibrations

[1] This feature was added subsequent to the research completed in conjunction with the original Gershon, 1992 study.

to the values obtained in the CUTHI/LO analysis. This run produced person ability measures based on the CUTHI/LO calibrations but included all item-person interactions.

In Condition 3, the item and person calibrations obtained in the Condition 1 analysis were used to edit the original data matrix. Using the CAT Software System (Gershon, 1992) we marked as missing all $-1 < B - D < 2$ (where B is the ability of the person and D is the difficulty of the item). The edited data set was then analyzed with BIGSTEPS with no additional persons or items deleted.

The difference between Condition 2 and Condition 3 is the point at which the data was edited. In Condition 2, the editing occurs after the PROX iterations. In Condition 3, data is edited after UCON estimates are obtained.

The output files generated from the BIGSTEPS runs were then graphed using the Item Characteristic Curve option in the CAT Software System (Gershon, 1992). The software program examines each person-item interaction using the person and item files generated by BIGSTEPS. In addition, it compares the answer key with the raw data file. For every item, the item difficulty is subtracted from the person ability. A tally is then kept for each quarter logit range on the B-D scale of the percent of items answered correctly. Theoretical, observed unedited, and observed edited results are plotted on the same scale. This plot compares the theoretical probability of correct response with the observed percent of correct response under each condition.

RESULTS

Figure 1 shows the BIGSTEPS (Wright and Linacre, 1992) Map of Persons and Items from the Condition 1 analysis. The map indicates that many examinees have greater than 90% probability of answering many of the items correctly. It further shows that this data set is appropriate for considering the CUTHI/LO option because the distance between person ability and item difficulty is extreme for many cases.

In Figure 2, the Condition 2 and Condition 3 item characteristic curves are shown. Figure 2 indicates that Condition 2 and Condition 3 produced identical results. This means that the item calibrations and person measures obtained from the BIGSTEPS CUTHI/LO analysis where the data was edited after the PROX iterations produced the same results as editing the data after UCON iterations. Since the results were effectively the same, and since using the CUTHI/LO option is by far the easier procedure, only the results for Condition 1 and Condition 2 will be further reported.

Impact of Deleting Item-Person Mismatches on the Item Characteristic Curve

Figure 3 and Table 1 show the comparison of the theoretical item characteristic curve (ICC) versus the unedited ICC and the ICC obtained from the CUTHI/LO analysis.

At the lower end of the scale where the difficulty of the item far exceeds the ability of the person, the Rasch model predicts less than 10% probability that a person will answer the item correctly. However, since these were multiple choice items with four distractors, one would presume a random probability of guessing an answer correctly closer to 25% of the time. A second discrepancy is observed where there is no difference between the ability of the person and the difficulty of the item. At this point, the Rasch model predicts that items are answered

correctly 50% of the time (Wright and Stone, 1979). The percent of examinees responding correctly in the unedited analysis was .448. When CUTHI/LO is employed, and only person-item interactions where $-1 <$ B-D < 2 are included, the percentage of examinees responding correctly was .479.

Impact of CUTHI/LO on estimated item difficulty

The impact of using CUTHI/LO on the estimated difficulty of items is seen in Figure 4. The correlation for the two sets of item calibrations is .99, but note that the standard deviations differ greatly. The standard deviation for the CUTHI/LO calibrations is 1.24 while the standard deviation for the unedited calibrations is 1.06. Figure 4 shows the difference between the identity line and the slope created by the ratio of the standard deviations (1.17). This difference is significant using an F test and comparing the ratio of the variances of the two sets of calibrations ($F_{113,113} = 1.37$, $p < .05$). Deleting inappropriate person-item interactions spreads the items difficulties when administered to an appropriate sample of examples; the easy items are easier and the difficult items are more difficult.

Impact of CUTHI/LO on estimated person ability

Figure 5 shows a comparison of the ICCs obtained from the Condition 1, unedited analysis, the Condition 2, CUTHI/LO analysis and the analysis when all items were anchored to the CUTHI/LO item calibrations. The later analysis anchored the item difficulty estimates obtained from CUTHI/LO, but *all* person-item encounters were included regardless of B-D. This analysis is comparable to what may typically be done in practice. After the "best" estimate for the item calibrations is obtained, item values are anchored and all persons and items are included in a subsequent analysis.

When the CUTHI/LO item calibrations were used but all person-item interactions were included in the table and graph, the percentage of examinees responding correctly was .458. This finding illustrates the reality of using the CUTHI/LO option when the goal of the analysis is to obtain person ability estimates. The estimates obtained using anchored item difficulty estimates are better than using an unedited data set. However, they are still subject to a lot of noise presumably caused by person misfit due to poorly targeted items.

Impact of CUTHI/LO on item fit

A comparison of the item mean squared infit statistic from the unedited analysis and the CUTHI/LO analysis (Figure 5) shows that fewer items misfit when CUTHI/LO is employed. Although there is no difference in the mean of the two statistics, a comparison of the log of the variances of the two analyses is once again significant ($F_{113,113} = 1.89$, $p < .05$). More items in the unedited analysis fit poorly and more of the items were likely to overfit.

Discussion

Most "traditional" paper and pencil based testing formats will produce maps similar to the one shown in Figure 1, as tests are often designed to allow persons to answer the vast majority of the items correctly. While this practice does little to improve the overall efficiency of the testing experience, there is much to be said for allowing examinees a primarily successful testing experience. Also, in the case of Minneapolis Public schools, the student population included Chapter 1, Special Education students and Limited English Proficient students. The range of estimated ability at Grade 1 was over 8 logits. Since no single test will be sufficient to examine such a heterogeneous group, the test is targeted to lower ability students. The impact of this type of test is to produce a data set in which many of the person-item interactions are

clearly inappropriate for accurate item difficulty parameter estimation.

Editing data sets is an integral part of good data analysis. Item estimation is improved when misfitting persons are removed from the data set and person estimation is improved when misfitting items are removed from the analysis. A basic tenet of the Rasch model is that good quality items have constant values for other so called parameters such as guessing and discrimination. This research confirms this to be true especially when appropriately targeted persons are included in the sample. Using CUTHI/LO allows for improved item estimation based on samples of examinees for whom the item is appropriately targeted even though the original (unedited) data might not otherwise be considered appropriate for this purpose.

The ability of the CUTHI/LO procedure to effectively take a mistargeted data set, and make it a good one has several advantages for test developers. Extreme items, which have otherwise been overly subject to either (a) the effects of guessing by low able examinees, or (b) the effects of sleeping by high able examinees who sometimes get easy items wrong, can now be safely included in an analysis. Large scale banking efforts are frequently stymied by poor pre-targeting of test items to appropriate ability samples. CUTHI/LO can be used to remove the effects of this poor targeting. Of course, this option cannot guarantee that enough data will be left to accurately estimate the item difficulty! But, the results of this analysis help us to understand that the "data" never really existed in the first place. On the other hand, the use of the remaining well targeted data that is associated with CUTHI/LO option may lead many psychometricians to re-examine thresholds for the amount of data needed for accurate parameter estimation.

The significant increase in the standard deviation of the items and the significant decrease in the standard deviation of the mean square fit statistic indicates that the CUTHI/LO procedure is producing improved item calibration estimates, leading to improved person ability estimates. While the "improvement" is relatively small for any single test, the effect is likely to be greatly magnified whenever equating is used. CUTHI/LO helps to release the constraints on the item difficulty range which are encountered when inappropriate person-item interactions are used to estimated item difficulty.

Editing person-item mismatches will be especially useful with the advent of computer adaptive testing. Given that pre-calibrated item banks (often calibrated with paper and pencil tests) are being used for adaptive testing, accurate item difficulty estimates are essential. Furthermore, since adaptive testing can take advantage of items across the person ability range, improved parameter estimation for "easy" and "hard" items becomes increasingly important.

Reference

Andrich, D. (1989) Constructing fundamental measurements in social psychology in J.A. Keats (Ed), *Mathematical and Theoretical Systems*, New York: Elsevier, p. 17-26.

Bergstrom, B., Gershon, R., and Brown, W. (May, 1993) *The Effect of Ability on Differential Item Functioning*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Brown, W., O'Gorman, K., Rogers, C. and Alm, C. (1993) *District-wide student achievement test results in spring, 1993*. Minneapolis: Research and Development Division in Public Schools.

Gershon, R. (1992a) *The effects of person-item mismatches on the integrity of the item characteristic curve*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Gershon, R. (1992b) *The Cat System Software* [software program]. Chicago: Computer Adaptive Technologies, Inc.

Gershon, R. (1990) *Rasch-model procedures used to build the JOCRF vocabulary item bank*. Technical Report 1990-3. Chicago: Johnson O'connor Research Foundation

Wright, B. and Linacre, M. (1992). *BIGSTEPS* [software program]. Chicago: Mesa Press.

Wright, B. and Stone, M. (1979). *Best Test Design*. Chicago: Mesa Press.

Figure 1

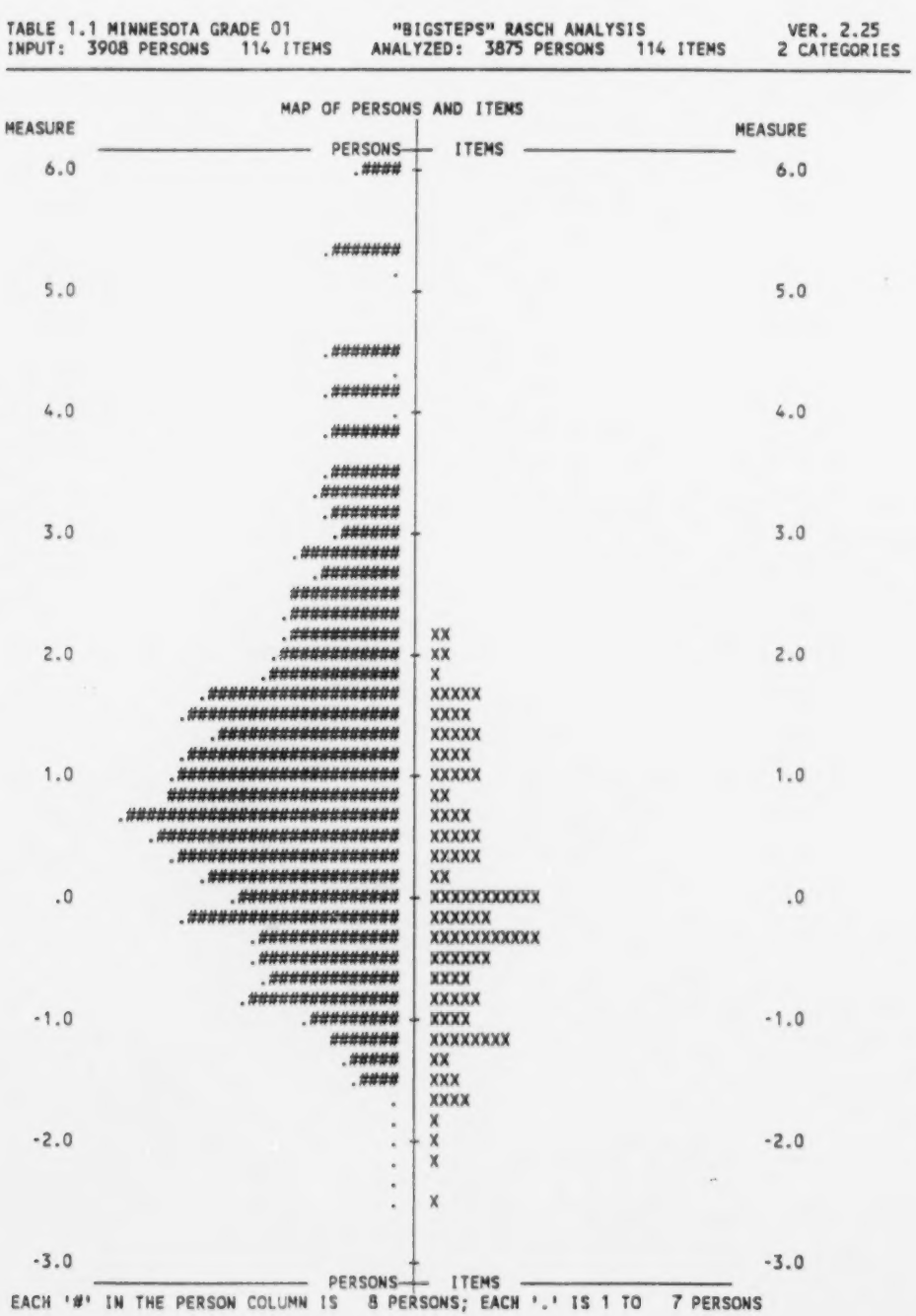


TABLE 1.1 MINNESOTA GRADE 01 "BIGSTEPS" RASCH ANALYSIS VER. 2.25
INPUT: 3908 PERSONS 114 ITEMS ANALYZED: 3875 PERSONS 114 ITEMS 2 CATEGORIES
MAP OF PERSONS AND ITEMS
MEASURE PERSONS ITEMS MEASURE
6.0 .#### 6.0
.#######
5.0 5.0
.#######
.#######
4.0 4.0
.#######
.#######
.########
.#######
3.0 .###### 3.0
.#########
.########
##########
.##########
.########## XX
2.0 .########## XX 2.0
.########## X
.################## XXXXX
.#################### XXXX
.################## XXXXX
.#################### XXXX
1.0 .################### XXXXX 1.0
################### XX
.###################### XXXX
.###################### XXXXX
.#################### XXXXX
.################### XX
.0 .############## XXXXXXXXXXX .0
.#################### XXXXXX
.############ XXXXXXXXXXX
.############ XXXXXX
.########### XXXX
.############## XXXXX
-1.0 .######## XXXX -1.0
####### XXXXXXXX
.##### XX
.#### XXX
. XXXX
. X
-2.0 . X -2.0
. X
.
. X
-3.0 -3.0
PERSONS ITEMS
EACH '#' IN THE PERSON COLUMN IS 8 PERSONS; EACH '.' IS 1 TO 7 PERSONS

Figure 2

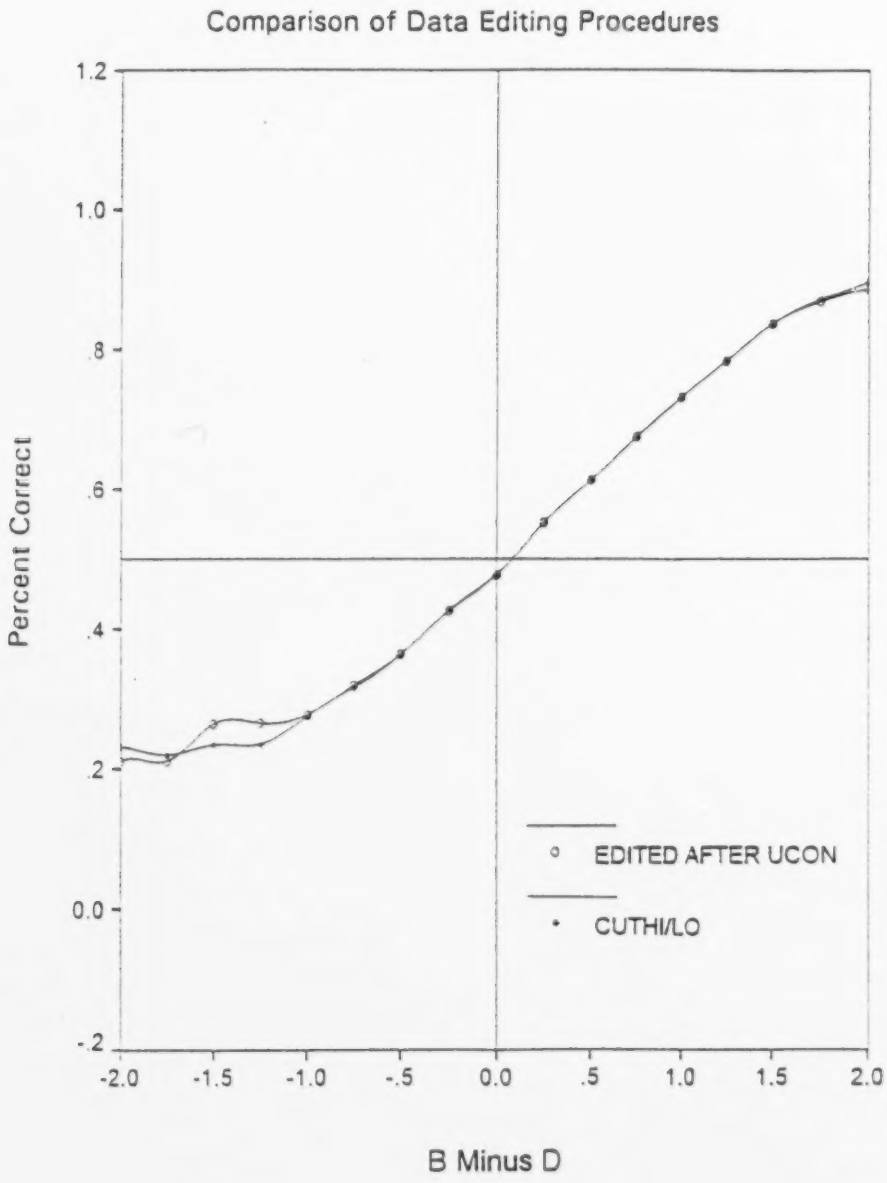

Comparison of Data Editing Procedures
Percent Correct
1.2
1.0
.8
.6
.4
.2
0.0
-.2
-2.0
-1.5
-1.0
-.5
0.0
.5
1.0
1.5
2.0
B Minus D
EDITED AFTER UCON
CUTHI/LO

Figure 3

Grade 1 Item Characteristic Curves

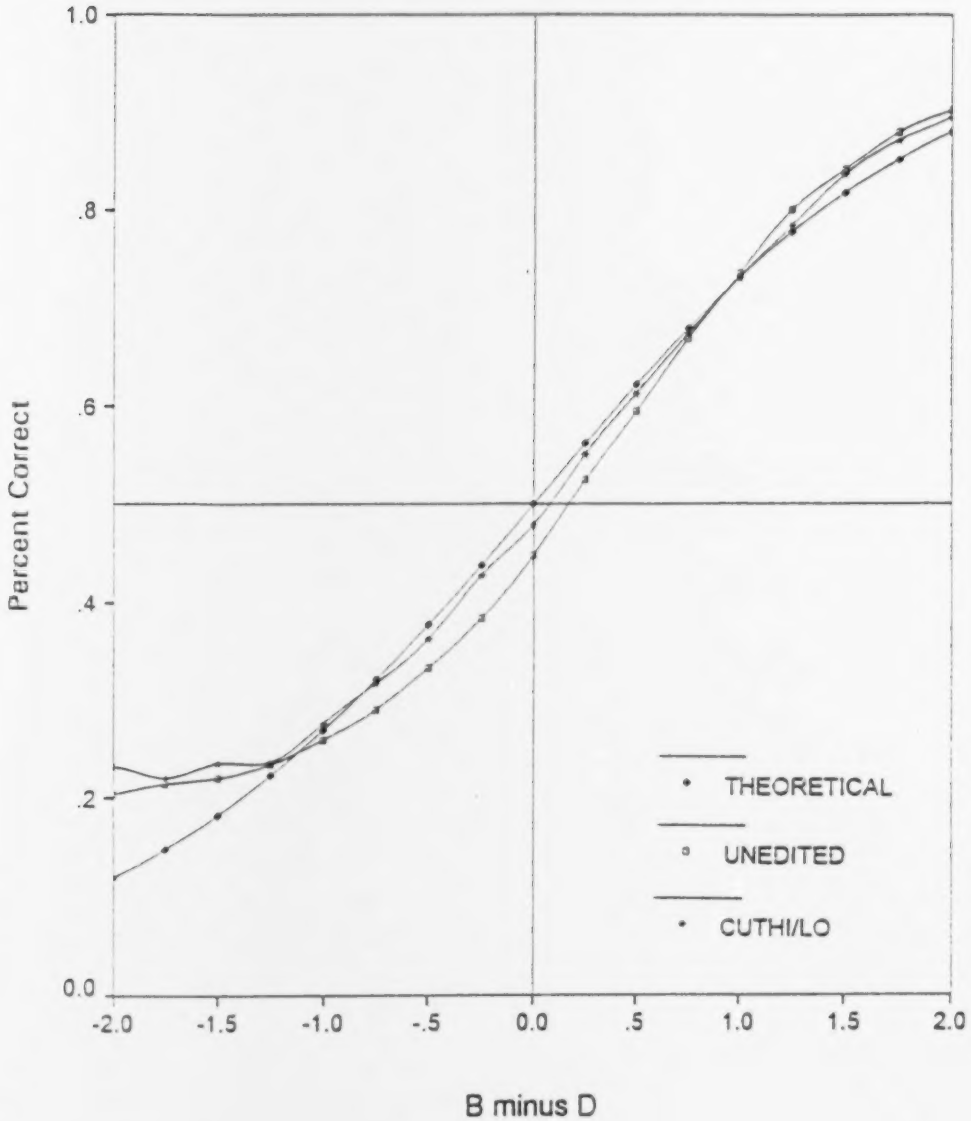
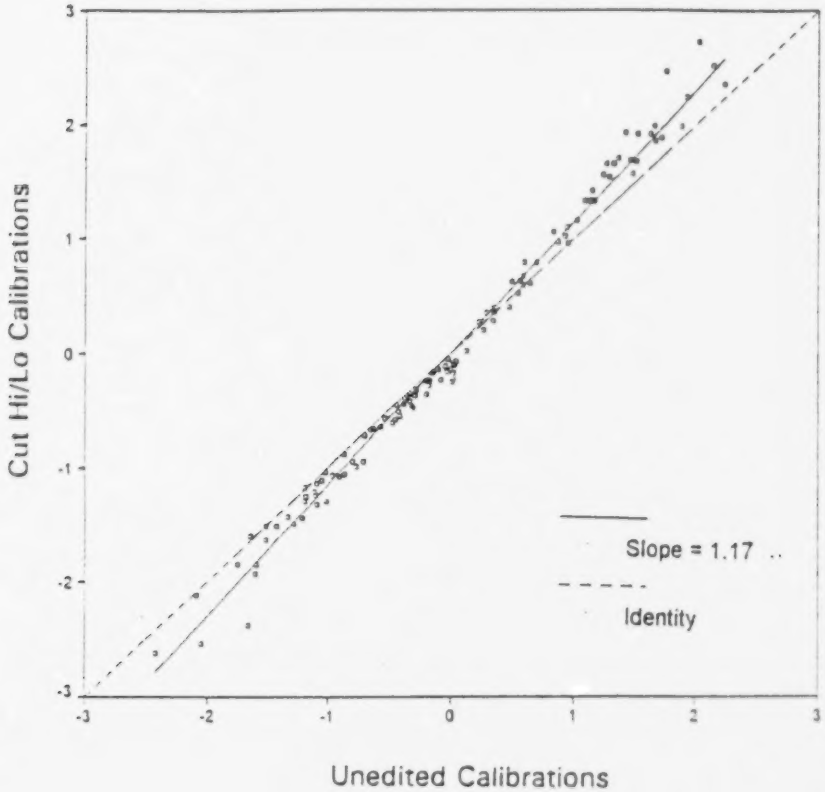


Figure 4

Unedited Calibrations vs CUTHI/LO Calibrations



UNEDITED
SUMMARY OF 114 MEASURED (NON-EXTREME) ITEMS

	SCORE	COUNT	MEASURE	ERROR	MNSQ	INFIT	MNSQ	OUTFIT
MEAN	2505.5	3718.8	.00	.04	.99	-.7	1.04	.7
S.D.	648.9	119.4	1.06	.01	.11	4.8	.31	4.4
RMSE	.04	ADJ.S.D.	1.06	ITEM SEP	24.08	ITEM SEP REL.		1.00

CUTHI/LO
SUMMARY OF 114 MEASURED (NON-EXTREME) ITEMS

	SCORE	COUNT	MEASURE	ERROR	MNSQ	INFIT	MNSQ	OUTFIT
MEAN	1338.0	2171.0	.00	.05	.99	-.3	1.00	-.2
S.D.	358.7	522.7	1.24	.01	.08	3.9	.12	4.0
RMSE	.05	ADJ.S.D.	1.23	ITEM SEP	23.58	ITEM SEP REL.		1.00

BEST COPY AVAILABLE

Figure 5

The Impact of Anchored Item Calibrations

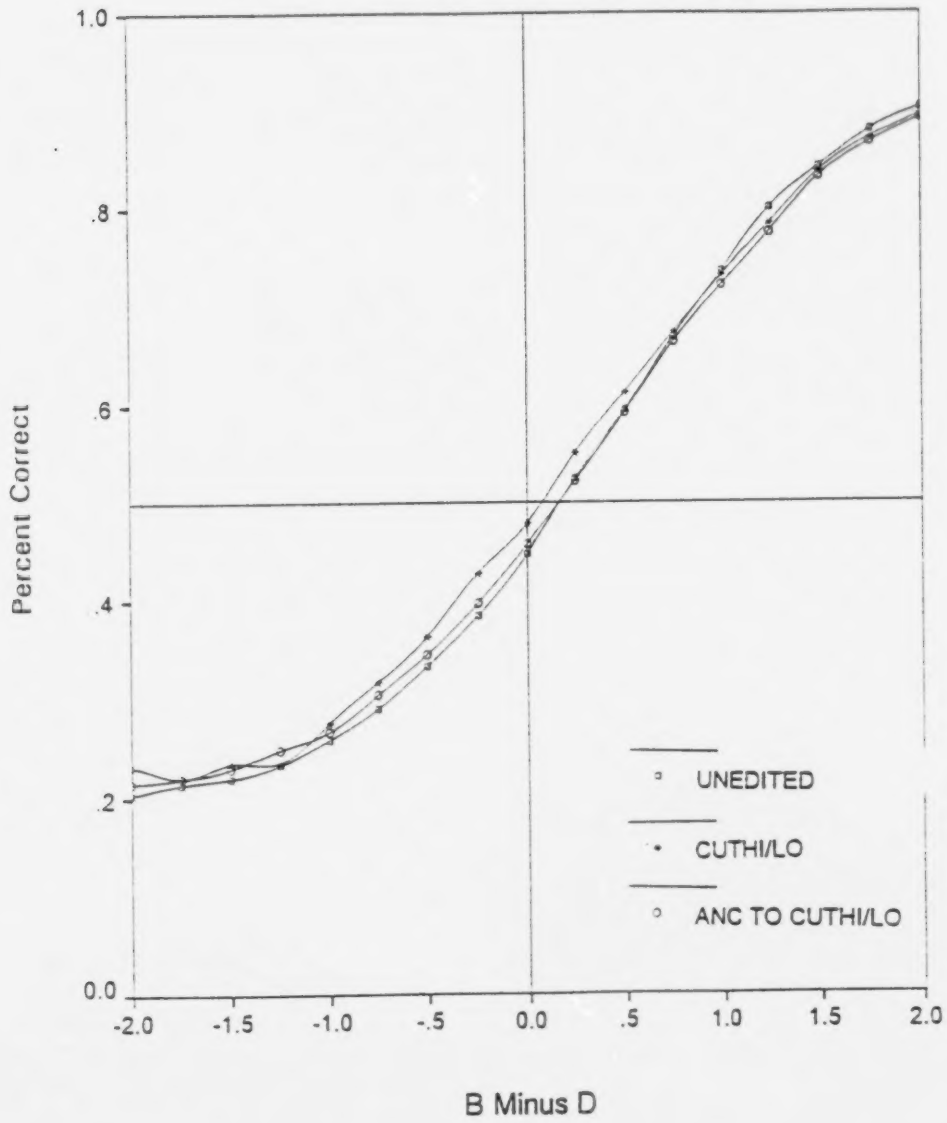


Figure 6

Comparison of Mean Square Infit

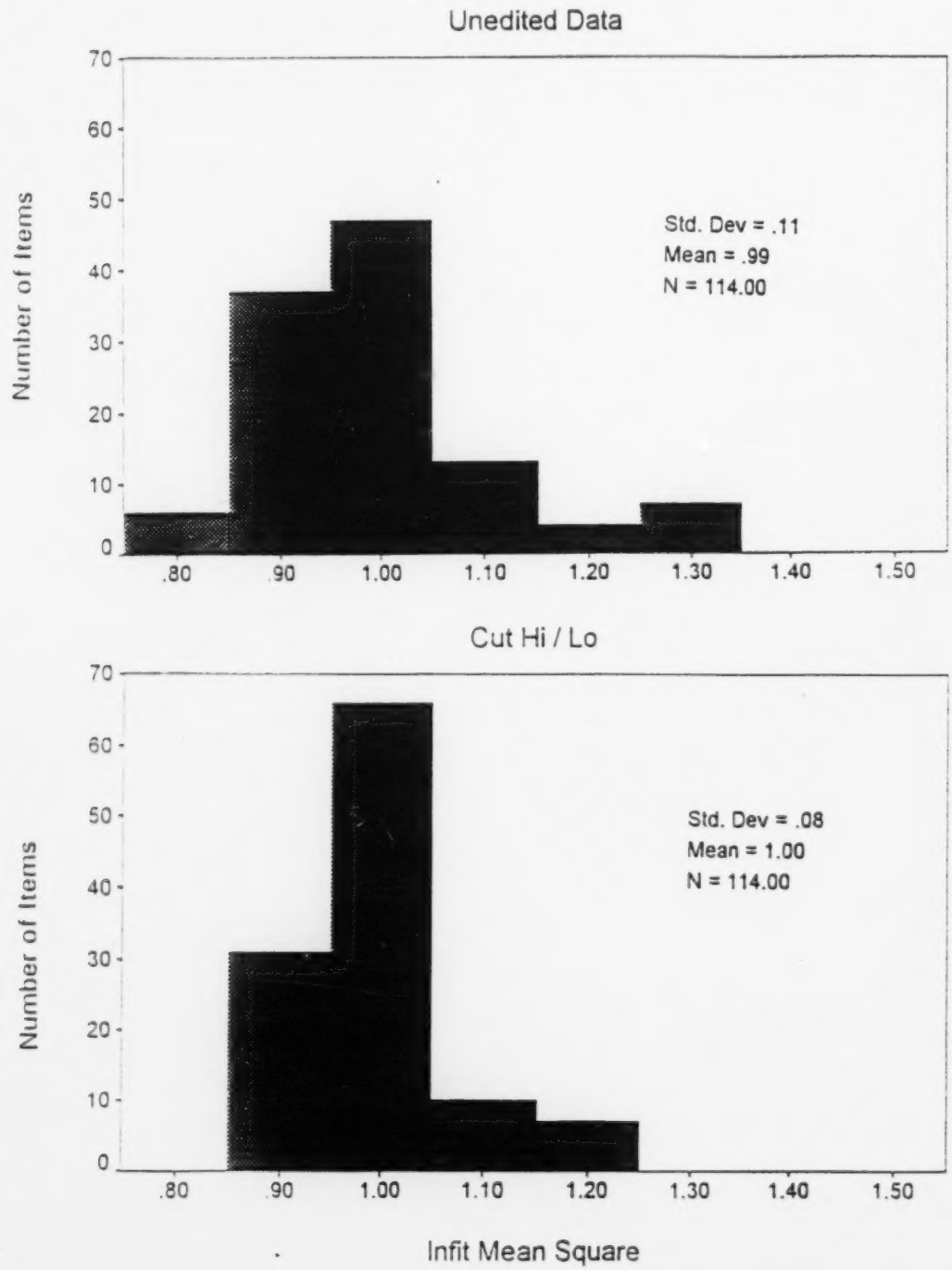


Table 1

B - D	THEORY	UNEDITED	CUTHI/LO	ITEM ANCHORED
-4.000	0.018	0.020	0.179	0.108
-3.750	0.023	0.059	0.194	0.171
-3.500	0.029	0.075	0.190	0.167
-3.250	0.037	0.133	0.193	0.183
-3.000	0.047	0.141	0.207	0.172
-2.750	0.060	0.174	0.225	0.184
-2.500	0.076	0.179	0.222	0.206
-2.250	0.095	0.194	0.222	0.212
-2.000	0.119	0.204	0.232	0.215
-1.750	0.148	0.214	0.220	0.220
-1.500	0.182	0.220	0.235	0.230
-1.250	0.223	0.234	0.236	0.249
-1.000	0.269	0.259	0.276	0.267
-0.750	0.321	0.290	0.317	0.304
-0.500	0.378	0.333	0.363	0.345
-0.250	0.438	0.384	0.428	0.397
0.000	0.500	0.448	0.479	0.458
0.250	0.562	0.525	0.551	0.522
0.500	0.622	0.595	0.613	0.592
0.750	0.679	0.669	0.674	0.664
1.000	0.731	0.735	0.731	0.721
1.250	0.777	0.800	0.784	0.775
1.500	0.818	0.842	0.837	0.832
1.750	0.852	0.880	0.872	0.867
2.000	0.881	0.903	0.896	0.892
2.250	0.905	0.921	0.923	0.911
2.500	0.924	0.936	0.930	0.924
2.750	0.940	0.946	0.938	0.935
3.000	0.953	0.959	0.941	0.947
3.250	0.963	0.963	0.945	0.957
3.500	0.971	0.967	0.954	0.964
3.750	0.977	0.975	0.958	0.966
4.000	0.982	0.988	0.964	0.982